Case Study 1: Predictive Data Analytics Identifying Loyal Customers: Organic Products Offerings (with Python)

Due date: 1st April, 2018
Weighting: 25%

Introduction

This assignment is intended to allow you to display your knowledge and understanding of predictive data mining. In this assignment, you will use classification algorithms (decision tree, regression, neural network and ensemble models) implemented in Python to display your technical competence gained from the practicals. It is also an opportunity for you to display the knowledge that you have gained from lectures and your readings and to show the relation between theory and practical.

The purpose of this assignment is to give you (1) an understanding that various methods can be applied to a data set and (2) the benefits of applying data mining techniques to a data domain.

Instructions

1. The assignment is due on 1tst April. It is a firm deadline.

2. You should submit the assignment via Blackboard Assignment.

3. The assignment (data mining results) **will also be marked in the practical class**. Each group member will be asked specific questions about the case study in **week 7** practical lab. A 15% marks (out of 25 marks) will be assigned to you on the individual performance.

4. This is a group assignment. It is your responsibility to form a team of 3 members and you should do so preferably by week 3. Groups are to be ARRANGED and MANAGED by you. As in real life, the performance of individuals in the team shall be judged by the performance of the team together, so choose your partners carefully.

5. To ensure that everyone agrees as to their responsibilities in the team and how you will work together, we have asked that you complete a Team Contract. This should be done before the team is registered. You can find the team agreement template and guidance under the Assessment Item 1 link.

6. Once the team is formed, complete the team contact and register the team on Blackboard. Choose “Tools” from the left side of panel. Select the “Groups” tool and choose one of the IFN645 groups to register. This should be done by week 3.

7. Of course, the work you (group) hand in must be your own; no collaboration or borrowing from other groups is permitted. We will use the usual methods of detection of any plagiarism.

8. All the datasets required for this assignment can be found in the provided file named as **casestudy1-data.zip**.

9. A report should be submitted via online submission answering each question of the case study. There is no need of including introduction, summary, conclusion or references in the report. The report should just include responses to the questions set in the case-study. Some answers may require screen shots. Use them as needed, but you may include your own table detailing those results. While you may like to go into extreme detail about, you will not have the space to do so. Rather, write down the important points and attach the important screen dumps, to show that you have thought the matter through. The report is expected to be about 15-20 pages long. Remember to include the final diagram of the project showing all nodes connected in your diagram.

10. Name the case-study report as **casestudy1.doc**. The word file should include a cover page with Student ID number and full name (as in QUT-Virtual) for all students, along with the group name. Combine this file with your **team contract**, and name the compressed file as **casestudy1.zip.** Submit this file on **Blackboard (under the Assignment 1 link)**.

11. Read the Assessment Policies on Blackboard or QUT Website.

Case Study Scenario

A supermarket has just begun offering a line of organic products. The supermarket has a customer loyalty program. As an initial buyer incentive plan, the supermarket provided coupons for the organic products to all of their loyalty program participants and has now collected data from 22,000 participants that includes whether or not these customers have purchased any of the organic products.

The supermarket's management would like to determine which customers are likely to purchase these products. You have been hired as a data analyst consultant by this management. Your task is to inform decision makers the (characteristics of) potential buyers from the entire user base by building predictive models on this data set of selective customers.

Case Study Dataset

The data set ORGANICS contains over 22,000 observations and 18 variables. The variables in the data set are listed in Table 1. The following information about variables would assist you in assigning the variables roles.

- The variables DOB, AGE, AGEGRP1, and AGEGRP2 are all different measurements for the same information.
- The variable NGROUP contains collapsed levels of the variable NEIGHBORHOOD.
- The variables LCDATE and LTIME represent the same information in two different formats.
- There are two target variables namely, ORGANICS and ORGYN, with different types. Choose the target that suits best to the given task.

Name	Description
CUSTID	Customer Loyalty Identification Number
GENDER	M = male, F = female, U = unknown
DOB	Date of birth
EDATE	Date extracted from the daily sales data base
AGE	Age, in years
AGEGRP1	Age group 1
AGEGRP2	Age group 2
TV_REG	Television Region
NGROUP	Neighborhood group
NEIGHBORHOOD	Type of residential neighborhood
LCDATE	Loyalty card application due
LTIME	Time as loyalty card member
ORGANICS	Number of organic products purchased
BILL	Total amount spent
REGION	Geographic region
CLASS	Customer loyalty status: tin, silver, gold or platinum
ORGYN	Organics purchased? 1 = Yes, 0 = No
AFFL	Affluence grade on a scale from 1 to 30

Table 1: List of Variables

Case Study Tasks

Your task is to build various predictive models such as decision tree, regression model, neural network and ensemble model on this data set and compare them. Results inferred by these models should inform decision makers the (characteristics of) potential buyers

Set up a new project for this task with **DMProj1** as the Python file and **ORGANICS** as the dataset. Include various models in this source file. Name all the models meaningfully.

Task 1. Data Selection and Distribution.

1. Can you identify any clear patterns by initial exploration of the data using histogram or box plot?
2. What is the proportion of individuals who purchased organic products?
3. Did you have to fix any data quality problems? Detail them.
4. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
5. What distribution scheme did you use? What data partitioning allocation did you set? Explain your selection.

Task 2. Predictive Modeling Using Decision Trees

1. Build a decision tree using the default setting. Examine the tree results and answer the followings:
    a. What is the classification accuracy on training and test datasets?
    b. What is the size of tree (i.e. number of nodes)?
    c. How many leaves are in the tree that is selected based on the validation data set?
    d. Which variable is used for the first split? What are the competing splits for this first split?
    e. What are the 5 important variables in building the tree?
    f. Report if you see any evidence of model overfitting.
    g. Did changing the default setting (i.e., only focus on changing the setting of the number of splits to create a node) help improving the model? Answer the above questions on the best performing tree.
2. Build another decision tree tuned with GridSearchCV. Examine the tree results.
    a. What is classification accuracy on training and test datasets?
    b. What are the parameters used? Explain your decision.
    c. What are the optimal parameters for this decision tree?
    d. What is the size of tree (i.e. number of nodes)? Is the size different from the tree built in the previous step (2.1)? Why?
    e. Which variable is used for the first split? What are the competing splits for this first split?
    f. What are the 5 important variables in building the tree?
    g. Report if you see any evidence of model overfitting.
3. What is the significant difference do you see between these two decision tree models (steps 2.1 & 2.2)? How do they compare performance-wise? Explain why those changes may have happened.
4. From the better model, can you identify which customers to target for further marketing? Can you provide some descriptive summary of those customers?

Task 3. Predictive Modeling Using Regression

1. In preparation for regression, apply transformation method(s) to the variable(s) that need it. List the variables that needed it.
2. Build a regression model using the default regression method with all inputs. Once you done it, build another one and tune it using GridSearchCV. Answer the followings:
    a. Name the regression function used.
    b. How much was the difference in performance of two models build, default and optimal?

c. Show the set parameters for the best model. What are the parameters used? Explain your decision. What are the optimal parameters?

d. Report which variables are included in the regression model.

e. Report the top-5 important variables (in the order) in the model.

f. What is classification accuracy on training and test datasets?

g. Report any sign of overfitting.

3. Build another regression model using the subset of inputs selected by RFE and selection by model methods. Answer the followings:

   a. Report which variables are included in the regression model.

   b. Report the top-5 important variables (in the order) in the model.

   c. What are the parameters used? Explain your decision. What are the optimal parameters? Which regression function is being used?

   d. Report any sign of overfitting.

   e. What is classification accuracy on training and test datasets?

4. Using the comparison statistics, which of the regression models appears to be better? Is there any difference between two models (i.e one with selected variables and another with all variables)? Explain why those changes may have happened.

5. From the better model, can you identify which customers to target? Can you provide some descriptive summary of those customers?

Task 4. Predictive Modeling Using Neural Networks

1. Build a Neural Network model using the default setting. After that, tune it with GridSearchCV. Answer the following:

   a. What are the parameters used? Explain your decision. What is the optimal network architecture?

   b. How many iterations are needed to train this network?

   c. Do you see any sign of over-fitting?

   d. Did the training process converge and resulted in the best model?

   e. What is classification accuracy on training and test datasets?

2. Refine this network by tuning it with GridSearchCV. Report the trained model, same as Task 4.1

3. Build another Neural Network model with inputs selected from RFE with regression (use the best model generated in Task 3) and selection with decision tree (use the best model from Task 2). Answer the following:

   a. Did feature selection help here? Any change in the network architecture? What inputs are being used as the network input?

   b. What is classification accuracy on training and test datasets? Is there any improvement in the outcome?

   c. How many iterations are now needed to train this network?

   d. Do you see any sign of over-fitting?

e. Did the training process converge and resulted in the best model?

f. Finally, see whether the change in network architecture can further improve the performance, use `GridSearchCV` to tune the network. Report if there was any improvement.

3. Using the comparison methods, which of the models (i.e one with selected variables and another with all variables) appears to be better?

   From the better model, can you identify which customers to target? Can you provide some descriptive summary of those customers? Is it easy to comprehend the performance of the best neural network model for decision making?

Task 5. Generating an Ensemble Model and Comparing Models

1. Generate an ensemble model to include the best regression model, best decision tree model, and best neural network model.
   a. Does the Ensemble model outperform the underlying models? Resonate your answer.
2. Use the comparison methods to compare the best decision tree model, the best regression model, the best neural network model and the ensemble model.
   a. Discuss the findings led by (a) ROC Chart and Index; (b) Accuracy Score; (c) Classification Report.
   b. Do all the models agree on the customers characteristics? How do they vary?

Task 6. Final Remarks: Decision Making

1. Finally, based on all models and analysis, is there a particular model you will use in decision making? Justify your choice.

2. Can you summarise positives and negatives of each predictive modelling method based on this analysis?

3. How the outcome of this study can be used by decision makers?

Marks Distribution

In data mining, there is hardly ever a single solution. The solution depends upon various setting such as input variables role and measurements, training size and the selected method parameters. You may find that your project partner may have different solution as yours. Your group should decide on a single project that you would like to be marked. Submit the report discussing the final project components.

We would mark your data mining project in the Week 7 practical class to explore your understanding of the data mining concept. You should be prepared to show your final diagrams and results panels to your marker. The marker will ask each individual student questions and will assign individual mark (~15%).

Assignment Components	**Marks**
Data Pre-processing	3.5
Decision Tree Models	5
Regression Models	6
Neural Network Models	5
Comparison: Predictive Models	2
Final Remarks: Decision Making	2.5
Team Agreement	1

Assignment 1 Criteria Sheet:

Criteria	Comments and scoring
Non Submission of all components/ evidence of plagiarism	0
Has demonstrated a task with a working model with /without submission and demonstrates the ability to run the program and add some components. Questions were poorly answered.	1-5
Has demonstrated a task with a working model having a data source, and diagram with the substantial but incorrect implementation of at least one of the three components (predictive models). Questions were poorly answered.	6-9
Has implemented models for all three tasks (three data mining algorithms) with at least one being substantially correct. Shows some understanding of concepts with some success applying knowledge in basic questions	10-13
Has implemented models for all three tasks: Two of the three tasks are fundamentally correct, with substantially correct work flow diagrams which may contain minor errors. Response to questions shows a fundamental understanding of terms and concepts.	14-17
Has fundamentally correct implementation of all five tasks i.e. correct allocations of a target, rejections of variables according to instructions, running three models and comparing them. Includes a demonstration of the competent application of tools. Almost all questions have been reasonably answered. Demonstrate a strong understanding of the methods and terms including predictive mining, partitioning, imputation, comparison node, ensemble, misclassification, average squared error, sensitivity, specificity, lift, ROC chart, lift chart, support and confidence during written analyses. Some minor errors are allowed. Written application is required to be of reasonable standard.	18-20
Has implemented all of the requirements above with very few errors. A strong focus on the application on creative application of tools, and evaluation and interpretation of results is evident.	21-23
All of the criteria above are met; extensive model generation and analysis have been conducted to produce exceptional outcomes and have applied principles learnt in lectures to enhance the results.	24-25